


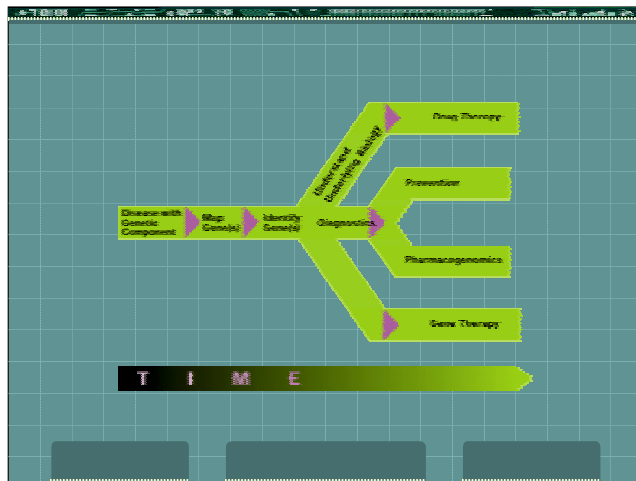
“I have always worked from the living model. I remember that once in the dissecting room when I was going over my ‘part’ with the demonstrator, he asked me what some nerve was and I did not know. He told me; whereupon I remonstrated, for it was in the wrong place. Nevertheless he insisted that it was the nerve I had in vain been looking for. I complained of the abnormality and he, smiling, said that in anatomy it was the normal that was uncommon. I was annoyed at the time, but that remark sank into my mind and since then it has been forced upon me that it was true of man as well as of anatomy. The normal is what you find but rarely.”

Introduction

- ♦ The new biology and the new generation
- ♦ Mountains of data and a new discipline
- ♦ Putting the new tools to work in therapy

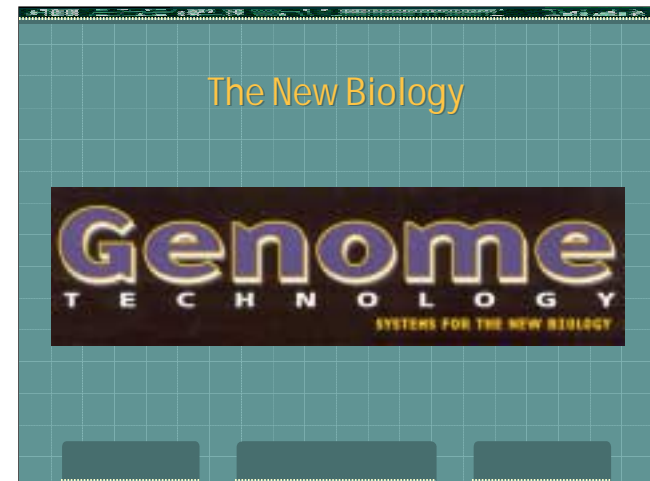


The Human Genome Project (HGP) began in the United States in 1990, when the NIH and the DOE joined forces with international partners to decipher the information contained in our genomes. The HGP began with a set of ambitious goals but has exceeded nearly all of its targets. Frequently ahead of schedule, HGP scientists have produced an increasingly detailed series of maps that help geneticists navigate through human DNA. They have mapped and sequenced the genomes of important experimental organisms. A working draft covering 90% of the genome in 2000. By 2003, we will finish the sequence with an accuracy greater than 99.99% - fewer than one mistake every 10,000 letters. HHMI 2000



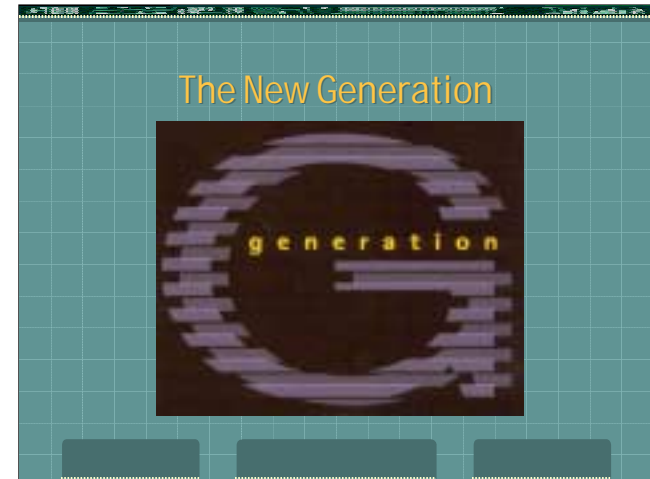
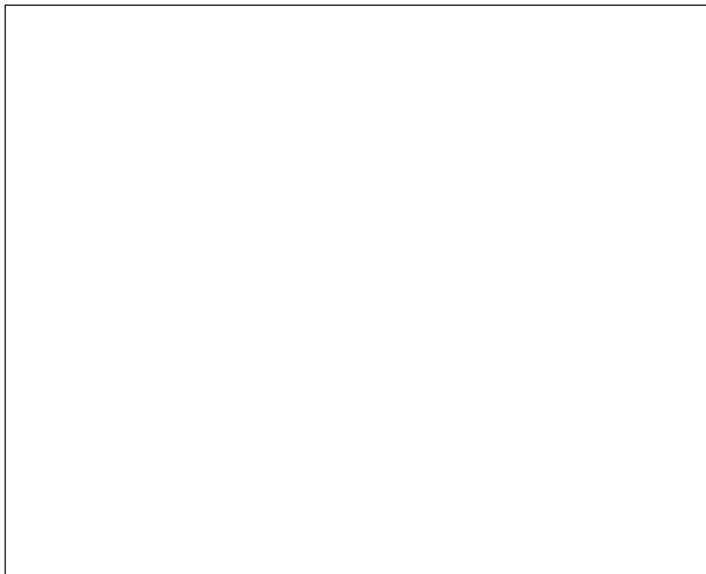
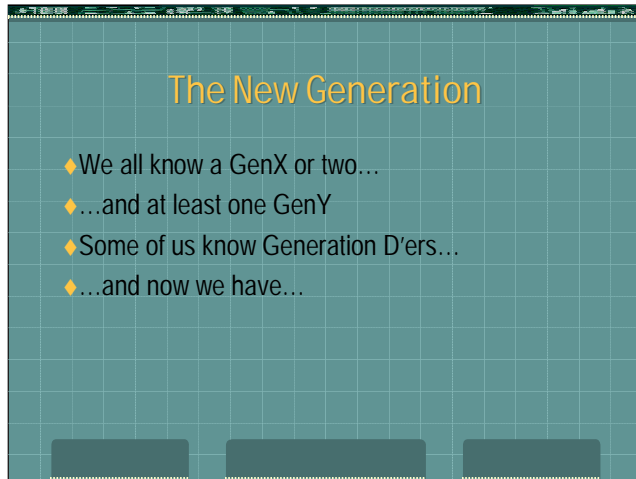
Molecular biology has long held out the promise of transforming medicine. Genomics will hasten the advance of molecular biology into the practice of medicine. As the molecular foundations of diseases become clearer, genetic tests will routinely predict individual susceptibility to disease. Diagnoses of many conditions will be much more thorough and specific. New drugs will target molecules logically. Drugs like those for cancer will routinely be matched to a patient's likely response. Diseases may be cured at the molecular level before they arise. All these changes aren't going to come quickly. It will take a long time to understand the human genome. But access to genome sequence will increasingly shape the practice of health care over the coming decades, as well as shed light on many of the mysteries of biology.

HHMI 2000



Genomics is is new biology and information technology supplies the systems for the new biology.

Genone Techonology 2001

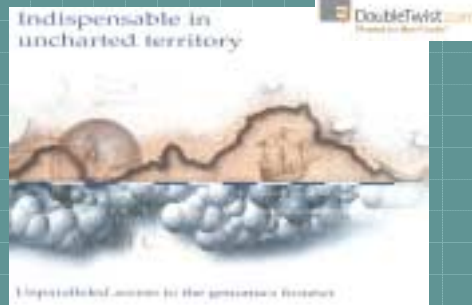


A recent issue of *Genome Technology* highlighted 16 rising genomic stars. When Watson and Crick described the structure of DNA none of the 16 had been born. When Lee Hood invented the automated sequencer, none of them had taken high-school biology. When the Human Genome Organization was established, many of them were home playing computer games. The same goes for the year Amgen was founded and the year Kary Mullis invented PCR.

They are 27 to 33. Most possess PhDs and have strong computer science skills. They are the G generation and they have the whole genome.

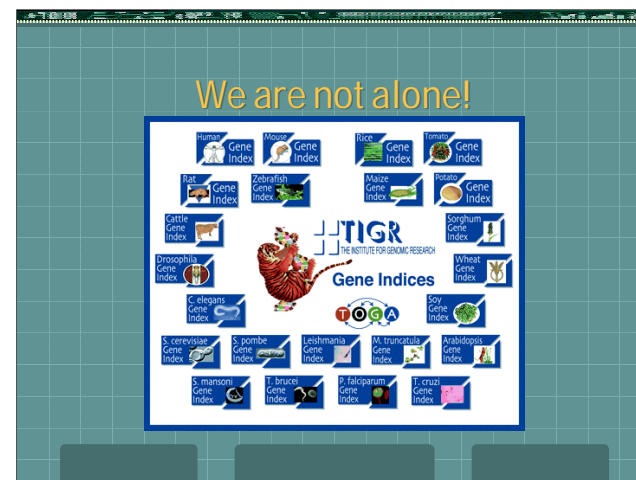
Genome Technology 2001

New Ads



New Companies





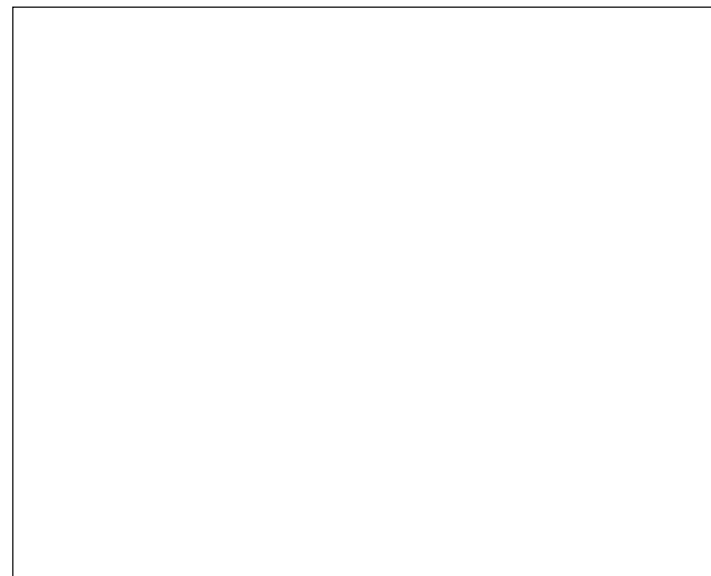
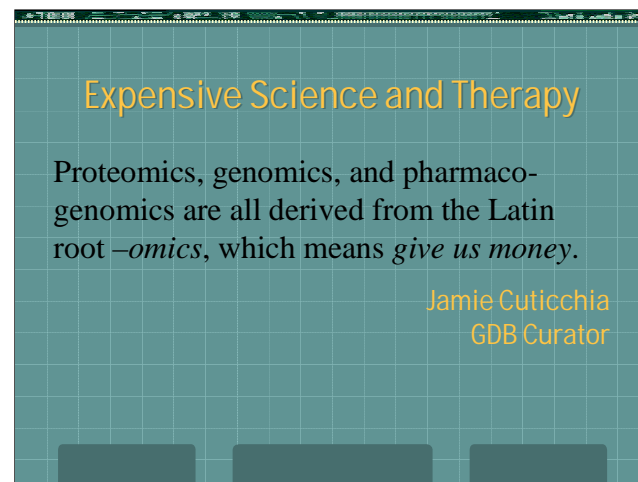
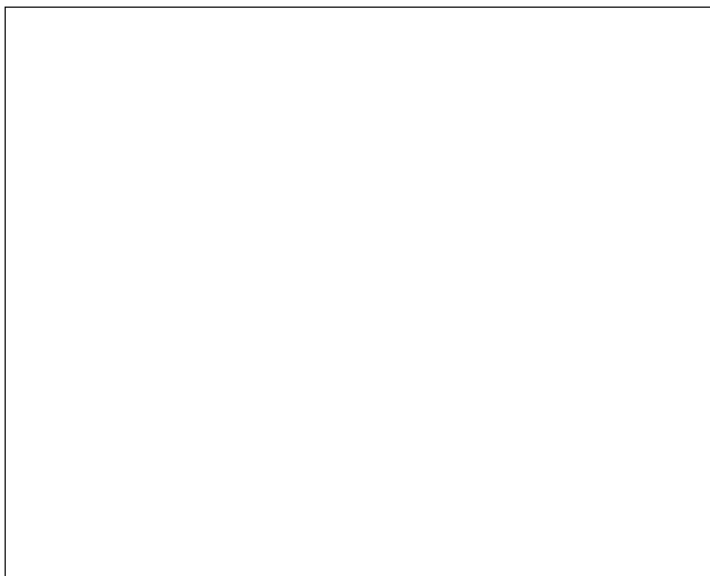


DOE and NIH collaborated to launch and nurture the Human Genome Project.



The Howard Hughes Medical Institute (HHMI) today unveiled a 10-year, \$500 million plan for a biomedical science center that will develop advanced technology for biomedical researchers and provide a collaborative setting where scientists from around the world can create the new tools of biology. The campus will be located on a 281-acre site that HHMI recently acquired just outside Washington, D.C., in Loudoun County, Virginia.

The Institute anticipates that the facilities on the new campus will be available for occupancy in about four years. The scientific staff will eventually number more than 200.



New Discipline

“Too much information running through my brain, too much information driving me insane.”

The Police
(a rock band)

New Discipline

Bioinformatics databases are some of the largest in the world. Celera, Inc’s database of human genome information is reportedly the largest private database in the United States.

Modern Drug Discovery
January 2001

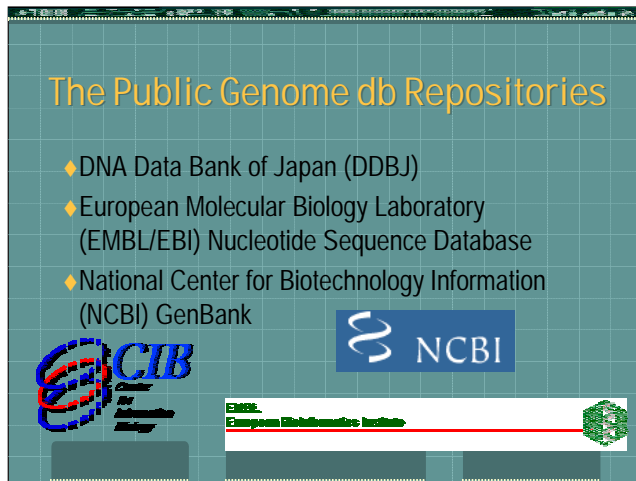
New Discipline

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

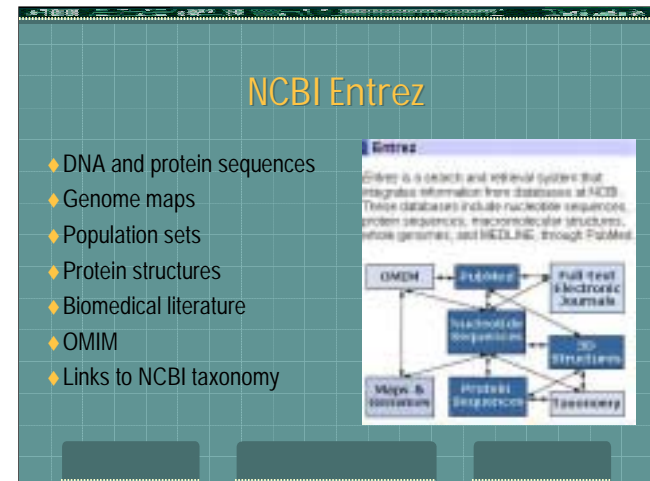
Although computer technology may be the father and biology the mother of bioinformatics, they both need to adjust to being parents after their successful single lives. Computer technology may have to return to its roots, when memory was dear and processors were less efficient, and biologists may have to learn more about computers to understand their child and teach the child prodigy to communicate. Even with parental compromises such as these, bioinformatics still must face its growing pains. Bioinformatics, when applied to genomics and proteomics, is required to compare and analyze multiple large databases simultaneously, something that few previous IT systems had to do. Modern Drug Discovery, Jan 2001

Who Uses Bioinformatics?





The Nucleotide database contains sequence data from GenBank, EMBL, and DDBJ, the members of the tripartite, international collaboration of sequence databases. EMBL is the European Molecular Biology Laboratory (EMBL) at Hinxton Hall, UK, DDBJ is the DNA Database of Japan (DDBJ) in Mishima, Japan. GenBank is the database maintained by the NCMI.




Entrez integrates the scientific literature, DNA and protein sequence databases, 3-D protein structure data, population study data sets, and assemblies of complete genomes into a tightly coupled system. The literature component of Entrez is known as PubMed.

The Nucleotide database contains sequence data from GenBank, EMBL, and DDBJ. Sequence data is also incorporated from the Genome Sequence Data Base (GSDB), Santa Fe, NM. Patent sequences are incorporated through arrangements with the U.S. Patent and Trademark Office (US PTO), and via the collaborating international databases from other international patent offices.

NCBI Entrez

- ◆ Text searches of sequence or bibliographic records
- ◆ Boolean queries
- ◆ Links to related information
 - ◆ Cross-references
 - ◆ Computed similarities
 - ◆ Sequences
 - ◆ MEDLINE abstracts



The diagram illustrates the Entrez system architecture. It shows a central 'Entrez' box at the top, which is described as a search and retrieval system that integrates information from various databases. Below this, several databases are shown in boxes, including OMIM, PubMed, Full-text Electronic Journals, Nucleotide Sequences, 3D Structures, Maps & Locations, Protein Sequences, and Taxonomy. Arrows indicate the flow of information and integration between these databases and the central Entrez system.

The Protein database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL and DDBJ as well as protein sequences submitted to PIR, SWISSPROT, PRF, Protein Data Bank (PDB) (sequences from solved structures).

The Genomes database provides views for a variety of genomes, complete chromosomes, contiged sequence maps, and integrated genetic and physical maps.

The OMIM database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at JHU and developed for the World Wide Web by NCBI. OMIM contains textual information and references. It also contains copious links to MEDLINE and sequence records in the Entrez system, and links to additional related resources.

NCBI Tools

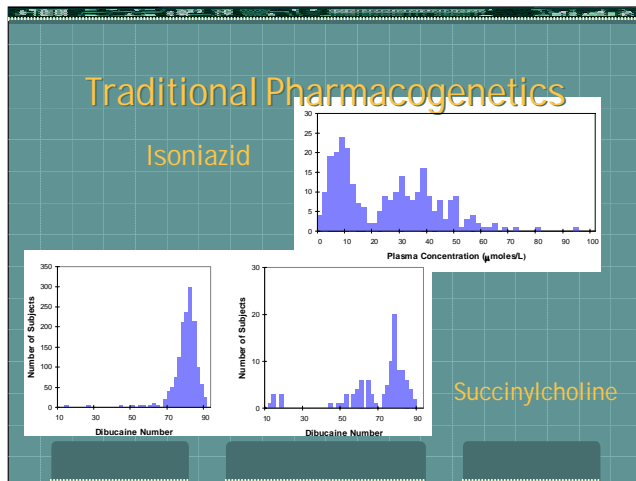
- ◆ DB retrieval tools
 - ◆ Entrez
 - ◆ The Taxonomy Browser
 - ◆ LocusLink
- ◆ BLAST family – sequence similarity search
- ◆ Gene-level sequences
 - ◆ UniGene
 - ◆ HomoloGene
 - ◆ RefSeq
 - ◆ dbSNP
 - ◆ ORF Finder
 - ◆ Electronic PCR
- ◆ Chromosomal sequences
 - ◆ Human Genome Map Viewer
 - ◆ Human Genome Sequencing
 - ◆ GeneMap '99
 - ◆ Human-Mouse Homology Maps
 - ◆ Cancer Chromosome Aberration Project (cCAP)
- ◆ Genome-scale analysis
 - ◆ Entrez Genomes
 - ◆ Clusters of Orthologous Groups (COGs)
 - ◆ Retroviral Genotyping Tools

NCBI Tools

- ◆ Gene expression & phenotype pattern analysis
 - ◆ Cancer Genome Anatomy Project (CGAP)
 - ◆ Gene Expression Omnibus (GEO)
 - ◆ SAGEmap
 - ◆ Online Mendelian Inheritance in Man (OMIM)
- ◆ Molecular structure
 - ◆ Conserved Domain Database (CDD)
 - ◆ Molecular Modeling Database (MMDB)

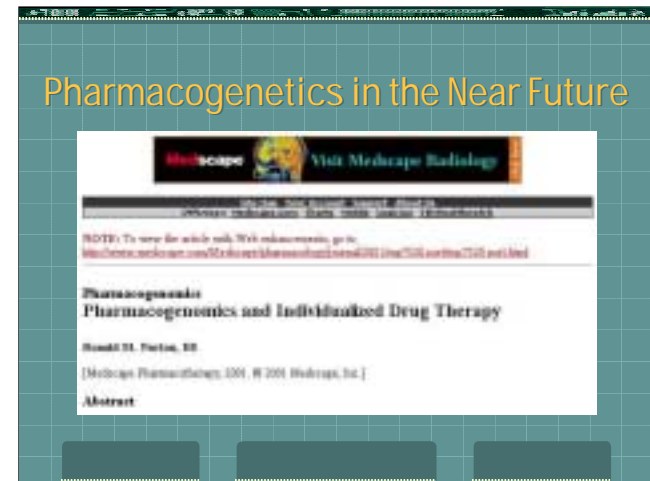
OMIM



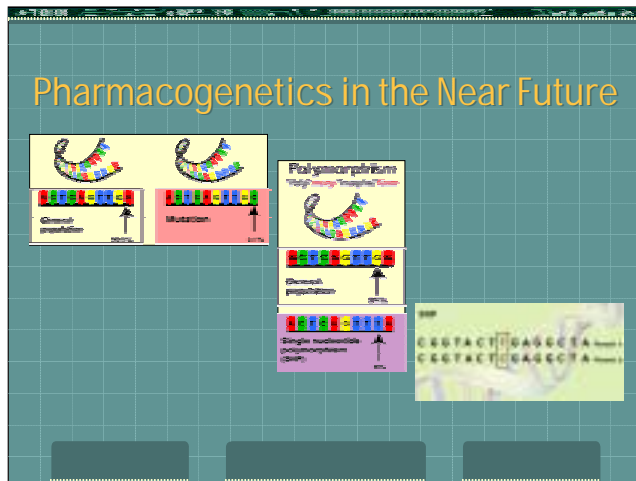


Traditionally pharmacogenetic problems have been identified retrospectively. Only after the fact do we learn there are populations that have a problem with the drug.

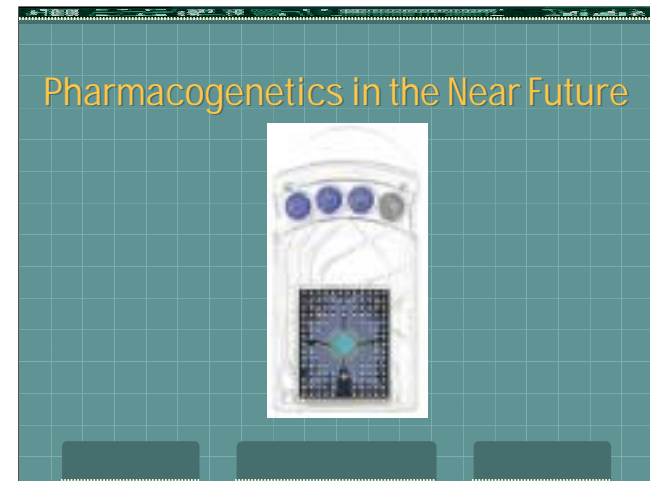
Plasma isoniazid concentration 6 hours after oral dosage at 71 mol/kg. Note the bimodal distribution for isoniazid. (From: Price-Evans, D. (1963). Am J Med **3**: 639- 662.) Distribution of plasma cholinesterase phenotypes in man. Dibucaine number is a measure of the percentage inhibition of plasma cholinesterase by 10^{-5} mol/L dibucaine. The abnormal enzyme has, in addition to low enzymatic activity, a low dibucaine number. (Left) Normal population. (Right) Families of subjects with low or intermediate dibucaine numbers. . (From: Kalow, W. (1962). Overview of Pharmacogenetics. New York, Pergamon Press.)



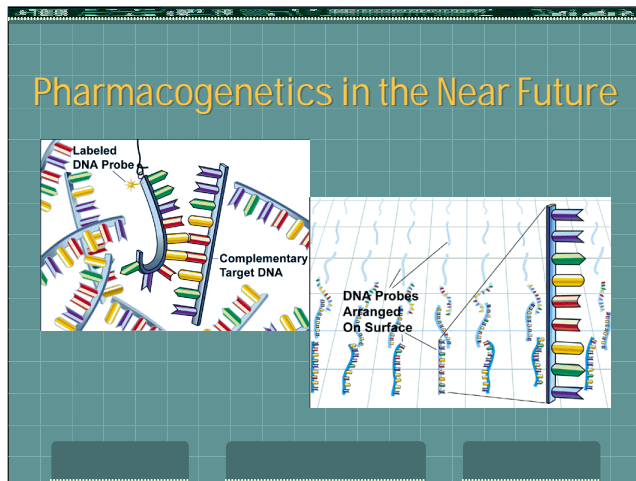
Pronounced “snip”, SNPs are single-nucleotide polymorphisms or one-letter variations in the DNA sequence. SNPs contribute to differences among individuals. The majority have no effect, others cause subtle differences in countless characteristics, like appearance, while some affect the risk for certain diseases and the response to some drugs. Pharmacogenomics describes the idea of tailoring drugs for patients, whose individual response can be predicted by genetic fingerprinting. Better understanding of genetics promises a future of precise, customized medical treatments.



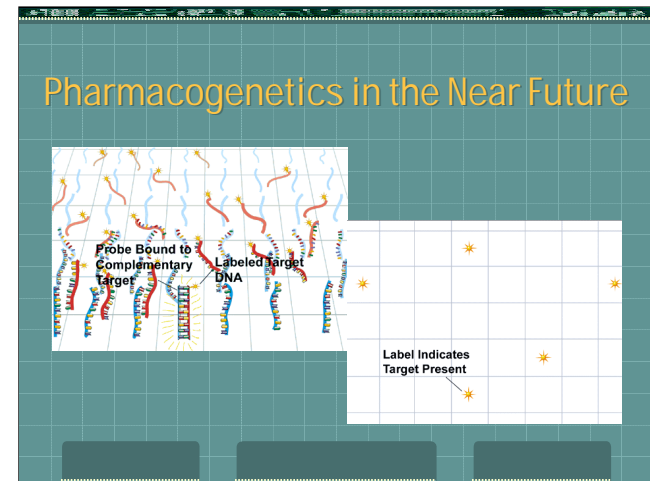
In *Science* a large group of researchers will describe a map of 1.42 million single nucleotide polymorphisms (SNPs) distributed throughout the human genome, providing an average density on available sequence of one SNP every 1.9 kb. These SNPs were primarily discovered through two large projects: The SNP Consortium and the Human Genome Project's analysis of clone overlaps. The map integrates all publicly available SNPs with described genes and other features of the genomic landscape. They estimate that 60,000 SNPs fall within exons, and 85% of exons are within 5kb of the nearest SNP. This high-density SNP map provides a public resource for defining haplotype variation across the genome, and should speed the identification of biomedically important genes as novel targets for diagnostic and therapeutic intervention. The SNP Consortium 2001



In DNA array technology, molecules are either synthesized on the chip, or pre-synthesized and then deposited. The DNA molecules range in size from short oligonucleotides (25mers) to cDNAs, to bacterial artificial chromosome (BAC) clones with inserts of 100 kilobases. These DNA probes can then be used to interrogate unknown target sequences based on specificity of hybridization to the known probes. Although this technology was first envisioned for application in DNA sequencing, the creativity of the community has been harnessed to broaden the utility of these chips for many aspects of functional genomics.

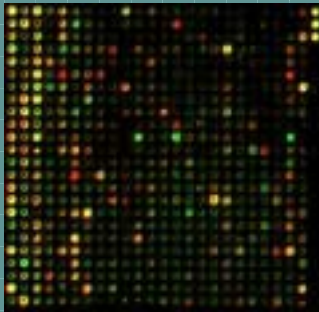


Scientists at Affymetrix Corporation and Stanford University developed two of the pioneering approaches in this field. In the Affymetrix approach high-density oligonucleotide arrays are synthesized on chips by photolithographic technology. For measurements of transcript expression a series of 20 oligonucleotides spanning the known sequence and 20 additional partner oligonucleotides with one-base mismatches to the target sequences are synthesized, hybridization is measured across the oligonucleotide population, and algorithms are employed to measure expression levels for each transcript.



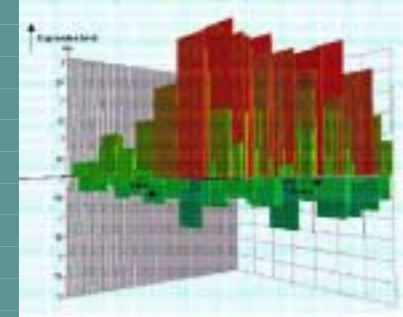
One of the great strengths of this approach is that sequence information in databases is used to design on-chip synthesis. Thus there is no need to maintain large collections of cloned DNA molecules. Additionally, because the oligonucleotides are relatively short and can be designed for any gene region, the technology can be applied to sequencing, identification of polymorphisms, and potentially for identification of different transcript splice variants. Affymetrix is now testing a chip that can detect 3,000 SNPs in less than 10 minutes. As the technology progresses, they expect to be able to mill through 100,000 SNPs dispersed through a patient's genome in several hours, for as little as a few hundred dollars.

Pharmacogenetics in the Near Future



A limiting factor for the technology is that, because photolithography is employed, the average laboratory cannot synthesize its own chips, and the use of photolithographic masks leads to relatively slow turnaround time for development of new chips. In the Stanford approach, cDNA molecules are spotted robotically on glass slides, and changes in gene expression are measured by labeling a control and experimental transcript population with different fluorescent tags and then measuring the intensity and ratios of the fluorescent signals. This approach has gained much popularity, in part because the technology has been disseminated widely, and the approach is well suited to rapid design and synthesis of new arrays with different sets of cDNA probes.

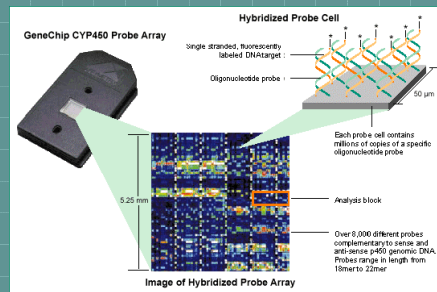
Pharmacogenetics in the Near Future



The power of the cDNA approach for gene expression analysis was convincingly demonstrated by Paul Spellman (Stanford University) in his very successful effort to catalog yeast genes whose expression is correlated with changes in the cell cycle. Using the power of yeast biology/genetics with the DNA arrays, Spellman identified more than 800 genes regulated in a cell cycle-dependent manner. (Strausberg, Robert L., and M. J. Finley Austin. Functional genomics: technological challenges and opportunities. *Physiol. Genomics* 1: 25–32, 1999.)

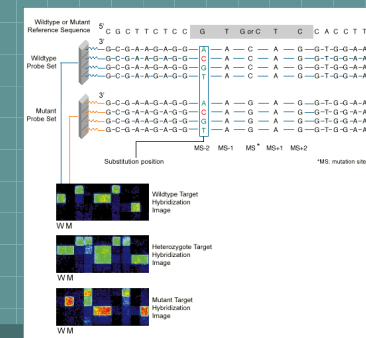
Example: mRNA was used to prepare fluorescently labeled cDNA (Cy3 for reference and Cy5 for succeeding time points). Yeast cell genes show a general increase in expression level during diauxic shift. (DeRisi *et al Science* 278:680-686.)

Pharmacogenetics in the Near Future



Development of expressed human enzymes, (human genes into yeast cells) makes it possible to test *in vitro* whether a given drug is metabolized by this particular enzyme. If it is metabolized, the biotransformation product and the enzyme kinetics can be determined. During drug development, compounds may be screened *in vitro* to determine if pharmacogenetic polymorphisms are likely to be involved in metabolism of the drug. Likewise, single-dose studies in subjects genotyped for various polymorphisms may clarify whether the potential for altered drug handling is clinically relevant. If a relatively rare, but severe, adverse reaction to a drug (e.g., a 1/5000 risk of hepatotoxicity) is strongly linked to a given pharmacogenetic polymorphism, such pharmacogenetic *pre-screening* could markedly decrease the risk for individual patients and the population as a whole.

Pharmacogenetics in the Near Future



For each known mutation site, there is a block containing 5 columns of probes complementary to wildtype sequence (W) and 5 columns of probes complementary to mutant sequence (M). The 5 columns interrogating wildtype sequence are interdigitated with the 5 columns interrogating the mutant sequence such that each pair of columns of probes interrogates the same nucleotide position in the target sequence.

Each pair of columns successively interrogates the target nucleotide sequence from two bases upstream of the mutation site to two bases downstream of the mutation site.

Each probe in a column contains a specific mismatch position called the substitution position where each of the 4 possible nucleotides are substituted into the probe sequence.

The Future is (almost) Now

- ◆ Herceptin and HER2 in 30% of breast cancer pts
- ◆ Codeine and CYP2D6 activation to morphine

Herceptin, shrinks some tumors and prolongs lives. What received much less attention, however, was the unique way in which Herceptin is prescribed. Herceptin is one of the first drugs for which tests are performed to predict whether it will work in a particular patient prior to drug prescription. The drug is specially designed to treat metastatic breast cancer patients whose tumors are shown to express abnormally high amounts of a protein called HER2. For those patients - up to 30 percent of women with breast cancer - Herceptin can bind to HER2, slowing tumor growth. The flip side: for those with normal HER2 levels, the drug is as useless a weapon as is a hilt without its blade. Hollon T, GeneLetter 1(12), January 2001

Information Resources

- ◆ Nucleic Acids Res - <http://nar.oupjournals.org/content/vol29/issue1/>
- ◆ NHGI - <http://www.nhgri.nih.gov/Data/>
- ◆ DOE - <http://www.ornl.gov/hgmis/education/education.html>
- ◆ HHMI - <http://www.nhgri.nih.gov/educationkit/>
- ◆ NCBI - <http://www.ncbi.nlm.nih.gov/Education/index.html>
- ◆ PhRMA - <http://genomics.phrma.org/>
- ◆ The SNP Consortium - <http://snp.cshl.org/>
- ◆ GeneChips - <http://www.gene-chips.com/>
- ◆ Bill's Bioinformatics Bookmarks - <http://www.toxicology.org/Education/CECourse/billsbookmarks.htm>